NOAA NESDIS
Cental Satellite Data Processing Center

# Comprehensive Large Array-data Stewardship System (CLASS) IT Architecture Description

July 20, 2001

Computer Sciences Corporation
Laurel, MD

# CLASS Executive Summary

**The goals set out for CLASS were:**

- Describe an architecture that can handle large array data sets, in particular those known to increase in volume
- Eliminate the need to keep creating "stovepipe" systems for each new type of data, but, in as much as possible use already polished sections of existing legacy systems
- Give any potential customer access to all NOAA (and possibly non-NOAA) data through a single portal
- Survey industry trends for processing systems, mass storage systems and storage management systems, to determine hardware (H/W) architectures and costs

To meet these goals the following CLASS architecture description is provided.

CLASS is a system that will unfold over a period of time. This is necessary to use a phased approach and keep cost, per year, to a minimum. The first phase of CLASS will be to design an access portal to all of NOAA's data, building a "toolkit" of items necessary for any data supplier (even non-NOAA) to use, and a configuration management discipline to follow to become accessible through the CLASS portal. This phase is an important phase as it allows the customers of CLASS to obtain the data they want, even though there are still separate "stovepipe" systems providing the data, through the single CLASS portal. The following phases will go on without the customer knowing about them, in other words, no loss of service, only an increase where possible.

The CLASS portal design will have minimum baseline requirements for a potential site to become a part of the CLASS network. The CLASS portal will require, at each site, a host server to house the portal software and prepare the data (sub-setting or super-setting). A relational database to house the CLASS inventory (catalog) metadata. A disk cache to stage the data, and a set of rules for indexing into the database, standardizing naming conventions, identifying back-up (mirrored) sites for particular data types, etc.

The CLASS portal's interface to the customer will be able to allow a novice to be lead through the various types of data they may be interested. It will also be able to aid the more advanced (intermediate) customer to cut through steps to get to the type of data they are looking for, and it will have even more short cuts for the advance customer that knows what they want. This interface needs to also allow for basic analyzation of data to aid the customer in choosing what to order.

The CLASS portal is the glue to CLASS but the CLASS is much more. CLASS will contain software that will properly interpret a query from the portal and be able to quickly discover the data type(s) that the customer is interested. This can be done through a networked relational database management system (RDBMS). The database(s) will identify all CLASS sites which contain the data (mirroring) in case of any network outages encountered when querying the closest site.

CLASS starts out with the portal into existing legacy systems and then phases in each of the legacy systems into the primary CLASS facility, as well as providing the tools and configuration management discipline for future systems to be built into the CLASS application. The goal for CLASS is to build one, object oriented, application software system which will be able to ingest, QA, archive and distribute all data types, adding one or more at a time.  This CLASS application will be maintained by one team, taking change requests from at least one backup CLASS site and providing updates in the form of version releases similar to the way vendors today provide COTS tools and version updates to the same, via CD or internet downloads.

Over time there will be a cost savings as maintenance of all of the “stovepipe” systems will dissolve into the one CLASS maintenance.  The fact that CLASS can access data from anywhere, through the portal, will allow NOAA management to physically locate the primary and backup CLASS sites where they will be the most economic.

# Table Of Contents

# 1 Introduction

The National Environmental Satellite, Data, and Information Service (NESDIS) is responsible for, among other things, the acquiring, archiving, and dissemination of environmental data collected by a variety of *in situ* and remote sensing observing systems from throughout NOAA and from a number of its partners [e.g., National Aeronautics and Space Administration (NASA)]. NESDIS has been acquiring this data for more than thirty years and will be archiving and disseminating even larger data sets in the future. Therefore, NESDIS is in need of an IT Architecture Description for a Comprehensive Large Array-data Stewardship System (CLASS) that will provide archive and access services for these data.

This IT Architecture Description Document is the final deliverable under Task Order 0019, from Department of Commerce (DOC)/National Oceanic and Atmospheric Administration (NOAA) to Computer Sciences Corporation (CSC) on the Central Satellite Data Processing (CSDP) contract (Contract Number: 50-SANE-6-00028).

# 2 Background

CSC, Short & Associates and Amdahl preformed an eight month task which consisted of reading and analyzing a volume of background material; interviewing individuals and groups from OSDPD, NASA, IPO, NGDC, NCDC, NODC and NCDDC; collecting information and metrics; analyzing the collected information; conducting an industry survey of processing systems, mass storage systems, storage management systems, long term storage media and telecommunications trends and costs; then compiling the totality of this information into the following IT Architecture Description.

# 3 Architecture Concepts

The architectural concept of the CLASS is straightforward. Utilizing state-of-the-art information technologies (IT), NESDIS will apply their existing sound, fundamental data management practices and procedures to satisfy two major mission requirements for large-array, long time series, remotely sensed data:

1) ***preservation*** of the entire science quality environmental data record; and

2) timely and efficient ***distribution*** of that data to customers.

CLASS will be a cooperative, coordinated rules-based architecture. It will be implemented through an evolutionary process of enhancing and integrating today's legacy systems in tandem with the design, development, and implementation of new systems and system components which are fully CLASS functional. A solid basis for implementation of a CLASS architecture can be found throughout existing NESDIS systems and system components, however none uniquely fulfill the full range of CLASS functionality. The NOAA Satellite Active Archive

(SAA) provides a good model for CLASS ingest-to-delivery functionality, yet lacks much of the data quality assessment and customer access processes and tools required. Ferret/Live Access Server (LAS) and SPIDR provide a good model for customer access and display tools, yet do not fully satisfy all of the CLASS access toolkit functionality required. The NOAA National Data Center (NNDC) Server satisfies much of the "common look and feel" CLASS requirements, yet lacks the integrated control and ease of system-to-system navigation required.

The structure of the CLASS architecture is based on a hierarchy of elements as shown in Figure 1. CLASS will be composed of independent, yet interdependent **sites** or **facilities**. Each CLASS facility will share a common set of functional **components**. These components form the basis of the major data stewardship roles provided by CLASS sites from ingest, through long term storage, to customer delivery. Finally, each component will be comprised of a series of **managed IT processes**. These processes will be defined under guidance and control of a CLASS-wide configuration management system, will be transportable from facility to facility, and will be designed as data independent activities. In the context of the CLASS architecture, the term *will* indicates a **requirement**, while the term *may* indicates an **option**.

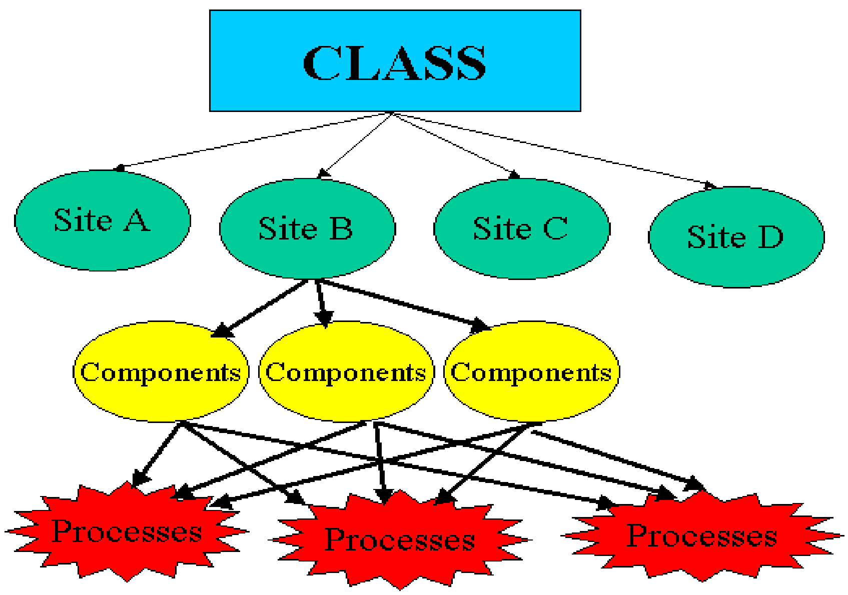


**Figure 1** CLASS Structure

## 3.1 Scope

NOAA Information Technology Architecture, Volume 1 Overview, June 2000 as well as the Gartner Group summarizes an IT Architecture as a framework and a set of principles, guidelines or rules to direct the process of acquiring, building, modifying, and interfacing IT resources throughout the Enterprise. These resources can include equipment, software, communications protocols, application development methodologies, database systems, modeling tools, IT governmental organizational structures and more. The purpose of this document is to establish such a framework for CLASS. Within the CLASS concept, every attempt has been made to identify the "what" without detailing specifics on the "how". Such specifics, rightly, are best left to the design phase of CLASS.

This document is presented as three groupings of related architectural elements:

- the functional components of CLASS
- conceptual views of CLASS from several perspective
- guidelines for design, development, implementation, and further enhancement of CLASS

## 3.2 Purpose

CLASS will be designed, developed, implemented, operated, and maintained for the primary purpose of supporting data and information stewardship for remotely-sensed environmental data. Initially, these data include observations from the following campaigns:

1) the National Oceanic and Atmospheric Administration (NOAA) and Department of Defense (DoD) Polar-orbiting Operational Environmental Satellites (POES);

2) the NOAA Geostationary-orbiting Operational Environmental Satellites (GOES);

3) the National Polar-orbiting Operational Environmental Satellite System (NPOESS);

4) the NPOESS Preparatory Program (NPP);

5) the National Aeronautics and Space Administration (NASA) Earth Observing System (EOS);

6) the NOAA NEXt generation weather RADAR (NEXRAD) Program; and

7) the European Meteorological Operational Satellite (METOP) Program.

Additionally, CLASS may encompass other large array data and data product systems including, but not limited to: NOMADS; NOS Side Scan, et. al.

GOES, POES, and NEXRAD are legacy data management systems within NESDIS and each will require significant enhancement and/or redesign to become part of the overall CLASS architecture. The remainder of these large array data systems will require initial design, development, and implementation as integrated portions of an existing CLASS architecture.

A second, though no less important goal for CLASS implementation will be the stewardship of all environmental data archived at the NOAA National Data Centers (NNDC), including smaller though no less important, *in-situ* data and older, no longer operational remotely-sensed data. These legacy data management systems may be integrated into the CLASS architecture as time, resources, and priorities dictate.

Finally, all new implementations of data management systems, including new *in-situ* and remotely sensed data streams for which NESDIS acquires archive and distribution responsibilities will be included in the CLASS architecture as implemented.

## 3.3 Goals

The goals of the CLASS architecture include:

- organizing information and its technologies to support NESDIS objectives of enhanced customer servicing;

- improving the effectiveness of IT in delivering new capabilities; and

- facilitating a continual evolution of IT infrastructure and solutions over time.

CLASS will be developed, implemented, maintained, and enhanced through the establishment of an enterprise technology environment that is based upon:

Interoperability: The architecture does not dictate that hardware technologies across the system be uniform, but rather that potentially disparate hardware implementations will operate in an integrated, interoperable environment. Interoperability will be considered at three levels of processing. Telecommunications interoperability will allow two facilities to exchange information transactions and share other forms of communication. Data interoperability will allow one facility to correctly and efficiently interpret data and information passed from another. Process interoperability will allow one facility to share the processes and services of another.

Extensibility: The architecture supports a plug-and-play approach through standardized process interfaces which allows for the incorporation of new IT resources as new requirements are determined.

Portability: The architecture provides the ability to move processes and applications from one facility to another as requirements, priorities, and resources dictate.

Flexibility: The architecture supports a variety of computing platforms, transitory data storage, and long term data storage. It strongly supports application portability. It allows for planned, incremental software and process changes and evolutionary introduction of new or enhanced technologies where applicable.

Affordability: The architecture supports the reuse of existing systems in as much as possible as well as the inclusion of Commercial Off The Shelf (COTS) software. Commercially available technologies support industry wide standards, provide periodic enhancements, and a growing platform independence.

Scalability: The architecture will be scalable in that sites need not be identical, but will provide the data stewardship functions within the scope and scale of the data they maintain. The architecture is adaptable, within the standards established for CLASS implementation, for both large array data processing facilities and facilities that process smaller, more complex data.

### 3.4 Assumptions

The following assumptions have been made in determining the CLASS target architecture:

- there **will** be two CLASS sites
- there **may** be more than two CLASS sites
- each CLASS site **will** have its data and information holdings replicated or mirrored at another CLASS site
- all of the Large Array data sets will be archived at CLASS sites
- three of the Large Array data sets are considered legacy systems (POES, GOES, and NEXRAD) and **will** either be migrated into CLASS through enhancement (POES/SAA) or redesign (GOES/GAA, NEXRAD/NAA?)
- the other Large Array data sets (MetOp, NPP, EOS, and NPOESS) **will** be designed, developed, and implemented at CLASS sites
- other NESDIS legacy (stovepipe) data management systems **may** be migrated to CLASS through enhancement or redesign
- other NESDIS, NOAA, and non-NOAA data management systems **may** be associated with CLASS by providing the functionality of access to data held within that system.

## 3.5 Structure

The initial topology of CLASS requires two physical sites separated from one another at a minimum of 50 miles to meet the NESDIS security requirement for disaster backup and restoration of data. From a customer view, these sites will appear as a single logical entity. Each site will, for all practical purposes, be constructed using identical hardware and software; will operate under the same configuration management guidelines; will maintain a dedicated high bandwidth telecommunications link between each other; will maintain a single, coordinated data and information catalog; and will operate as 24x7 data management facilities. Differences between the sites will occur in the specific data holdings and in the processes within the system which are data type dependent. The specific data holdings at each site are to be determined. Each site will operate as a mirrored site to the other, i.e., data and information holdings at one site will be replicated at the other.

There are two major architectural schemes for CLASS. The first encompasses what will be referred to as a full CLASS implementation, or **kernel architecture**. It supports all of the functional components of CLASS (see Section 4) and requires major resource expenditures for implementation, enhancement, operations, and maintenance. The kernel system is the framework for data stewardship of all large-scale NESDIS data streams. It will serve as the target architecture for NESDIS legacy data management systems.

The second scheme encompasses what will be referred to as a **CLASS associated architecture**. It satisfies two overall implementation requirements: it provides a method for integrating today's legacy data management systems into the CLASS arena with minimal resource expenditure; and it allows systems which will never be fully integrated into CLASS to satisfy a certain portion of CLASS functional requirements, primarily in the area of customer services. Over time, it is envisioned that many non-NOAA data management systems may become CLASS compliant.

### 3.5.1 Kernel Architecture

Initially, CLASS will be designed with a minimum baseline architecture that each CLASS site (primary and at least one backup) must possess to satisfy the minimum functional requirements for any one or a suite of data sets, i.e., two backup CLASS facilities may share the data set load or the primary site. The architecture will be **modular** in that it can expand to meet future processing needs, i.e., if an organization takes on the task of archiving and distributing EOS data, there should be a known hardware/software/ telecommunications expansion capability within the CLASS architecture to support the implementation of the additional workload and the integration of the new data stream into the overall CLASS framework of customer services. It will be **transportable** in that additional sites anywhere on the globe may be installed *at the minimum level* or at an expanded level to meet CLASS data management requirements. It will be **managed** in that all processes will be controlled by a single, integrated configuration management system based on a set of CLASS principles.

#### 3.5.1.1 Hardware

The baseline **hardware** architecture at each CLASS facility will consist of:

computer processor(s);

on-line, transitory storage devices;

near-line/off-line, long-term storage devices;

telecommunications devices; and

other peripheral devices necessary to support the entire operation and maintenance of the overall configuration.

This hardware, potentially varied in scope and scale initially, will maintain certain consistencies in implementation. In particular, all devices must support the open systems concept. All equipment will be operated and maintained, though not necessarily manned, in a 24x7 mode.

#### 3.5.1.2 Software

The baseline **software** architecture at each site will include, but not be limited to:

an operating system capable of integrating all hardware components;

an operating system capable of providing multiple, manageable processing environments or environmental partitions (e.g., MVS, UNIX);

state-of-the art telecommunications support systems, including at a minimum TCP/IP protocols and file transfer protocol (FTP);

programming languages capable of supporting object oriented design, development, testing, integration, and operation (e.g., Java, C++, etc.); and

a relational database management system (RDBMS); and

a common, low-level Applications Process Interface (API) to direct and control distributed access to the RDBMS (e.g., JDBC from Sun Microsystems, Inc.).

#### 3.5.1.3 Networking

The baseline networking architecture at each site will include:

high bandwidth connectivity to major suppliers of data, in particular those from POES, GOES, NEXRAD, EOS, NPP, NPOESS, and MetOp;

high bandwidth connectivity to Internet hubs; and

integrated internal pathways among clustered processors and storage devices.

#### 3.5.1.4 Design Guidelines

Multiple CLASS facilities will be integrated through the implementation of a CLASS-wide Configuration Management (CM) environment. Within this environment, technology refreshment and software development, testing, integration, and operation will be controlled. This CM environment will be instrumental in reducing overall cost of operations through the use of common, integrated technologies where appropriate; through elimination of redundancies in software development and implementation; and through implementation of standardized practices and procedures for managing and enhancing IT architectures.

The hardware architecture (processors, storage facilities, and telecommunications infrastructure) at the several sites will be functionally similar, though not necessarily identical. Though possibly disparate, the architectures will all rely on a common set of rules (CM); a common base of process management founded upon the principles of object-oriented (OO) system design, development, and implementation; and a common set of environmental data stewardship goals established by NOAA/NESDIS management and extensible through agreement with other environmental data collection, archival, and distribution stakeholders. The sites will conform to a to-be-established methodology for sharing developmental activities, upgrading technologies, and providing shared resources for processing integration, fail-over operations, and system security. This methodology will be established and maintained initially by the NESDIS Information Technology Advisory Team (ITAT).

The key to a successful implementation of CLASS will be the overall CLASS process management . In particular, CLASS will be designed and implemented to perform as a **single, logical entity**. To this end, the traditional concept of middleware (catalogs, inventories, system control scripts, etc.) will be expanded to include NESDIS-wide standards and goals for:

client applications;
presentation management applications;
business function management applications;
system integration applications; and
resource management applications

### 3.5.1.5 CLASS Catalog/Inventory

A critical feature of the CLASS kernel architecture is the CLASS catalog. This catalog will be an integral tool for the successful implementation of CLASS in that it will maintain *all* of the information required for distributed functional components (see Section 4) of the several CLASS facilities to interact with one another. The catalog will provide, but not be limited to inventory information about data sets, data set descriptions, other metadata pertaining to data quality and quantity, pointers to physical data set locations, and other quantifying and qualifying data set information. The catalog will play a fundamental role in providing a CLASS system that is seamless in appearance to the customer, in rapidly accessing data and information across the several CLASS facilities, and in standardizing the processes required to manage data within CLASS.

The elements of the catalog will be distributed across CLASS. Each facility will implement, operate, and maintain only that portion of the overall catalog which is associated with their data holdings. Internal processes (see Section 4) that require interaction with data and information will operate through the local catalog. Processes which require interaction with disaster backup facilities will interact directly with the appropriate remote catalog element(s). CLASS-wide processes, in particular customer access tools and techniques will interact with the distributed catalog elements across the Internet as if these distributed elements were physically part of the local catalog.

The catalog will operate with:

physically distributed and mirrored relational database management system(s);

structured using OO design techniques and technologies;

availability across the Internet, but within the CLASS security firewall;

Java based applications, rather than SQL based;

accessible via COTS APIs *such as* Sun's JDBC;

CM rules and guidelines for content, structure, and

24x7 availability guaranteed.

### 3.5.2 CLASS Associated Architecture

An alternate architecture is focused on providing access to legacy data management systems and systems which may never become full partners within CLASS. This approach will be useful during any transition of existing systems to CLASS and for the inclusion of smaller systems, possibly those maintaining only one or two small data sets for which a full scale CLASS implementation would be impractical. Under this approach non-CLASS systems may provide a **baseline data access and distribution functionality of CLASS** through implementation of this architecture. This architecture is not intended to replace the CLASS kernel architecture as an approach for data stewardship of large array data sets.

This approach provides data management services that would be in the form of a one-way access to data and information, not the two-way approach defined for the CLASS kernel systems. Full CLASS facilities and CLASS customers would have access to data and data products held within the external system, however customers of the external system would have no *direct* access to the CLASS archives and customer services. The overriding benefit of CLASS-compliant architecture would be the implementation of a transition architecture for legacy/stovepipe systems. An additional benefit would be an expanded distribution of site specific data.

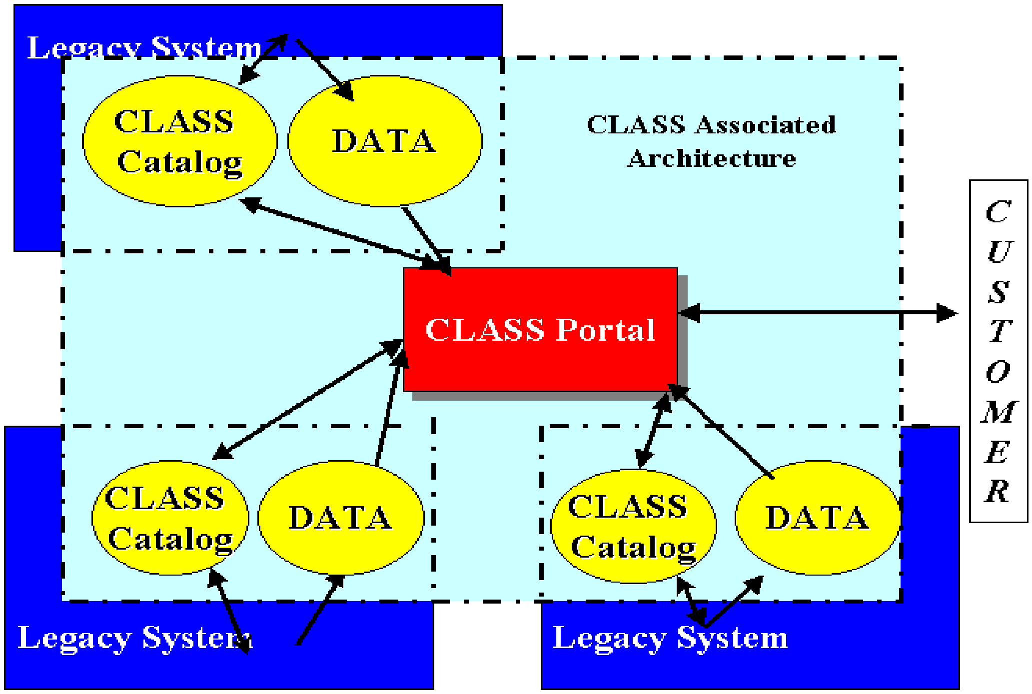


**Figure 2** CLASS Associated Architecture

The following components and functionality at non-CLASS sites would satisfy requirements for CLASS data and information distribution: see fig. 2 above.

- maintenance of one or more data sets appropriate to CLASS data stewardship;

- inclusion of full data set descriptions and documentation in the CLASS catalog, using CLASS standards and procedures;

- maintenance of a local portion of the distributed CLASS catalog;

- provision of Data Retrieval and Data Delivery components that would respond to process requests from any CLASS Data Retrieval component;

- provision and maintenance of a "trusted" network connection between CLASS sites and the external system;

- provision of computing facilities sufficient to process CLASS data and information requests, including catalog access;

- provision of transitory, on-line storage for staging CLASS-requested data sets; and

- compliance with CLASS configuration management practices and procedures for all components that interact with CLASS sites.

CLASS associated sites/systems would not require (from a CLASS perspective) that data be stored on long-term media, nor that local Data Ingest , Data Storage, or Customer Access functional components follow CLASS policies and procedures. They would not require that the *totality* of their Data Storage and Data Retrieval components comply with CLASS, only those portions that interact with CLASS. CLASS associated sites/systems would not have guaranteed disaster backup and restore facilities offered through CLASS nor share in the design, development, testing, implementation, and maintenance of the CLASS kernel.

The overriding benefit of the CLASS associated architecture would be the implementation of a transition architecture for legacy/stovepipe systems. An additional benefit would be an expanded distribution of site specific data.

# 4 Functional Components of CLASS

CLASS will be based on a series of interconnected and interoperable data and information management components, each of which will satisfy a major functional data stewardship requirement area. These functional components will be organized within an envelope of a CLASS process management configuration and the entire CLASS architecture organized and integrated through an overall system management configuration. These components will be designed, developed, tested, integrated, operated, and maintained under configuration management (CM) processes and procedures as a series of transportable software, procedures (scripts, naming conventions, OO compositions, etc.), and operating “rules” packages. The packages will be delineated as major and minor releases and intermediate “bug fixes” as the situation warrants.

The functional CLASS data management components will include:

- Data Ingest
- Data Preparation and Quality Assessment
- Data Storage
- Customer Access
- Data Retrieval/Repackaging
- Data Distribution

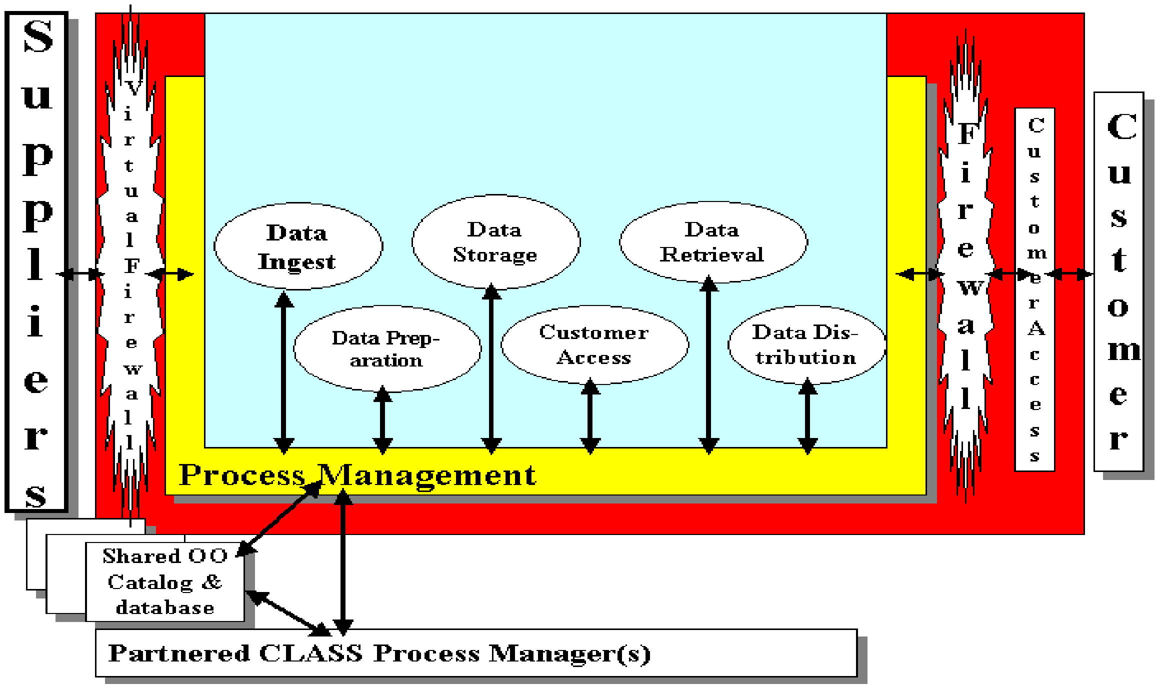


**Figure 3 CLASS Functional Overview**

Of these components, Customer Access will be the key to implementing an architecture that goes beyond the more traditional role of data management systems where data preservation provides the primary focus. The primary focus of CLASS will be shared between archive **management** and the ability of customers to locate, access, merge, and acquire data from multiple observation systems and multiple environmental science disciplines to meet their particular criteria. In essence, CLASS will significantly improve the usefulness of data and information held within it.

## 4.1 Data Ingest

The CLASS ingest facility will be an all day every day (24x7) operation. The operation will be time critical for remote-sensed operational data streams as data and information are ingested and processed at a near-real-time rate. Ingest targets are such that data complete the ingest process with *n* hours of observation where *n* is established through mutual agreement between NESDIS and the data supplier. The facility may be manned or unmanned at any point in time, however the operation will be automatically and continuously monitored such that any component (hardware/software/telecommunications) or data problems which require human intervention have appropriate maintenance personnel available. Prolonged system downtime will result in data loss; negative impacts on the data provider(s); and/or severe impacts on processing and storage facilities, thus hardware, storage, and telecommunications redundancies, “hot” backups**,** and fail-over operations will be provided.

### 4.1.1 Process Characteristics

| Process | Description |
|---|---|
| Receipt | Data will be ingested as made available by the provider through either backplane connectivity (co-location), through dedicated, NESDIS managed high speed telecommunication links, via relatively low speed Internet services, or via computer compatible media. |
| Initial Storage | Data will be hosted on network attached or network accessible digital media. Through agreement with provider(s) data will be maintained at the provider facility until such time as the CLASS ingest process is completed. |
| Data Assessment | Data transmissions or deliveries will be automatically tested for completeness (sent = received), integrity, and replication (whole or in-part) with previous transmissions. Retransmission may be required, reverting to Process 1, above.<br><br>Optionally, data will be automatically assessed for content or scientific consistency. For time critical data streams, this quality assessment will be performed within a targeted ingest process end-to-end time**.** |
| **Inventory** | Initial inventory information will be automatically derived from data set naming convention or a combination of naming convention, data scan and possibly determined data quality. |
| Subscription Preparation | Initial inventory information will be scanned for predefined customer selection criteria for delivery. Data which meet these criteria ("push" or "pull") will be migrated (physically or virtually) to customer-accessible storage facilities. |
| Subscription Delivery | Predetermined set of customers ("push" subscribers) will have data meeting their selection criteria delivered via dedicated telecommunications links, via the Internet, or on computer compatible media |
| Subscription Notification | Predetermined set of customers ("pull" subscribers) will be notified of data availability within their selection criteria. |

| | |
|---|---|
| Data Depiction | Certain data sets will be automatically scanned and various data depictions (browse, summaries, etc.) created. |
| Preparation  Notification | The processes which perform archive preparation and quality assessment  functions will be automatically notified when data are available for continued processing. |
| Provider Notification | The **data** provider will be automatically notified of completion of the CLASS ingest function for each data set. |

### 4.1.2 Architecture Characteristics

| Technology Category | Requirements |
|---|---|
| Telecommunications | Link to operational data provider(s) will be via backplane connectivity or via dedicated, high-speed, NESDIS managed link(s).<br><br>Links to other electronic data providers will be via dedicated links or via the Internet.<br><br>Link to electronic subscription customers will be via backplane; dedicated high-speed links; and/or the Internet |
| **Processors** | High speed, dedicated processors with automated redundancy (hot backups) will be directly gated or routed to telecommunications hardware for operational data .<br><br>Dedicated or shared processors will be directly gated or routed to high-speed telecommunications hardware and/or address routed to Internet gateways. |
| On-line Storage | Dedicated, high density digital storage with automated redundancy/duplication; fail-over* operations; and matched to a projection of incoming data volume.<br><br>Dedicated and shared high density digital data storage matched to projected subscription demand. |
| **Long term Storage** | System and periodic data disaster recovery storage system for backup and restore processing |

| | |
|---|---|
| Application Software | -inventory interface<br>-data set evaluations<br>-optional quality assurance testing<br>-subscription inventory interface<br>-automated subscriber notification<br>-data depictions |
| Process Management Interface | Relational Data Base Management System (RDBMS) for data inventory and subscriber management control.<br><br>RDBMS for disaster backup and restore management. |

* Fail-over is defined to be automated selection of redundant files by the operating system.

### 4.1.3 Architecture Interfaces

| Component | Requirement |
|---|---|
| External Architecture Data Delivery | Process interactions with all external (to CLASS) electronic data providers, including protocol handshakes, two-way information transactions (messaging), and data throughput. |
| Data Processing/Quality Assurance | two-way information transactions |
| Data Delivery | one-way (to) information transactions |
| **Legacy System Data Delivery** | Process interactions including two-way information transactions (messaging), and data throughput |

## 4.2 Data Preparation/Quality Assessment

The CLASS data preparation facility provides for both the automated and interactive quality assessment of ingested data and either triggers the re-ingest of data sets or provides for the internal modification of ingested data sets to comply with NESDIS data management standards. These functions cannot be bypassed in the normal sequence of processing. For data sets subject to re-ingest there will be established, automated procedures with data suppliers to accomplish this function. Either a re-ingest or data set modification will require automated notification of action to both providers and subscribers.

### 4.2.1 Functional Characteristics

| Process | Description |
|---|---|
| Automated and Interactive Data Assessment | Various algorithms may be applied to ingested data sets, including generation of data depictions, summaries, etc. useful in determining scientific and technical quality. |
| Interactive Data Modification | NESDIS scientists and/or technicians may interactively modify data sets or metadata to reflect appropriate changes and/or corrections. |
| Inventory | Data set inventories will be automatically updated noting any change in the quality, quantity, and or status of each data set, and provide a historical traceability. |
| Re-Ingest Notification | For electronically ingested data the provider will be automatically notified of the necessity to re-ingest data sets when necessary. |
| Subscriber Notification | Subscribers of all or a portion of each data set modified or re-ingested will be notified of such action and the reason(s). |
| Storage Notification | The processes which perform long-term, data storage functions will be automatically notified when data are available for continued processing. |

### 4.2.2 Architecture Characteristics

| Technology Category | Requirements |
|---|---|
| Telecommunications | No addition to Data Ingest. |
| Processors | High speed, dedicated processors with automated redundancy (hot backups) will be directly gated or routed to telecommunications hardware |
| On-line Storage | No addition to Data Ingest. |
| Long term Storage | No addition to Data Ingest. |
| Application Software | -inventory interface<br>-automated provider notification<br>-data depictions |

| | |
|---|---|
| Process Management Interface | Storage Management System (SMS) for long term storage process<br><br>RDBMS |

#### 4.2.3 Architecture Interfaces

| Component | Requirement |
|---|---|
| Data Ingest | one-way (from) information transactions |
| Data Storage | one-way (to) information transactions |
| External Architecture Data Delivery | one-way (to) information transactions |

## 4.3 Data Storage

The CLASS storage facility is a 24x7 operation. It will provide for long-term, near-line storage of ingested data sets. Additionally, each physical facility will operate in partnership with at least one other CLASS facility to provide mutual disaster backup/recovery services. All long term storage may be robotically controlled; will be transparently accessible to all applications software; and may be managed by COTS data management systems as an integral part of the facility process management system.

#### 4.3.1 Functional Characteristics

| Process | Description |
|---|---|
| Physical Storage | Data will be copied from network attached or network accessible (on-line) storage media to near-line, long term archive storage. |
| Remote Disaster Backup Storage | Data will be automatically copied from network attached or network accessible (on-line) storage media *to* a remote CLASS site for disaster backup/recovery protection. As an intermediate step, the remote site may cache data to on-line digital media before processing to long term storage. Process management information is exchanged automatically between the two sites. |
| Inventory | Data set inventories will be automatically updated. |

| | |
|---|---|
| Local Disaster Backup Storage | Data will be copied *from* a remote CLASS site for which the facility operates as the disaster backup. As an intermediate step, the facility may cache data to on-line digital media before processing to long term storage. Process management information will be exchanged automatically between the two sites. |

### 4.3.2 Architecture Characteristics

| Technology Category | Requirements |
|---|---|
| Telecommunications | Link to backup site via dedicated, high-speed, NESDIS managed link(s) matching projected one-way traffic volume.<br><br>Link to any site for which local backup services via dedicated, high-speed, NESDIS managed link(s) matching projected one-way traffic volume for each remote site. |
| Processors | No addition to Data Ingest |
| On-line Storage | No addition to Data Ingest for facility data sets.<br><br>If serving as a backup for a remote site, dedicated, high density digital storage with automated redundancy/duplication; fail over operations; and matched to projection of incoming data volume. |
| Long term Storage | High-density, rapid access, system-transparent digital media equivalent to archive period-of-record-to-date plus a minimum of 150% of the anticipated annual volume ingested.<br><br>If serving as a backup for a remote site, high-density, rapid access, system-transparent digital media equivalent to archive period-of-record-to-date plus an anticipated annual volume ingested. |
| Application Software | -inventory interface<br>-media management interface |

| | |
|---|---|
| Process Management Interface | SMS<br><br>Matched RDBMS and robotic facility management software to remote site if serving as a disaster backup. |

### 4.3.3 Architecture Interfaces

| Component | Requirement |
|---|---|
| Data Preparation/Quality Assessment | one-way (from) information transactions |
| Data Storage (CLASS remote facility) | two-way information transactions and two-way data throughput |

## 4.4 Customer Access

The CLASS Customer Access component is a 24x7 operation. The CLASS customer access portal will allow for data discovery of any data set held in the CLASS system, regardless of physical location. It will allow for some basic analysis of the data to ensure it meets the customer criteria. It will also contain a registration and billing and accounting function.

### 4.4.1 Functional Characteristics

| Process | Description |
|---|---|
| Present Information | Information regarding the resources available at CLASS sites will be provided to the customer including: basic system information; options for payments; a CLASS Web Customer Toolkit; the results of customer queries; etc. This functionality will be provided in a non-secure (outside the firewall) area. |
| Gather Information | Information regarding customer direction for CLASS processing will be gathered including search criteria; customer profile; payment information; help requests, etc. This functionality will be provided in a non-secure (outside the firewall) area. |

| | |
|---|---|
| Parse Requests | Process management information will be parsed into requests and forwarded to the appropriate CLASS Data Retrieval and/or Data Delivery processes. This functionality passes information through the firewall to secure CLASS areas |
| Parse Responses | Process management information will be parsed into results and forwarded to customers from the appropriate CLASS Data Retrieval and/or Data Delivery processes. This functionality passes information through the firewall from secure CLASS areas |
| Web interface | A graphical user interface (GUI) with the capability to search the CLASS inventory for data discovery. It will provide customer access to inventories of archived data sets, the ability to depict/display browse images that exist for the associated data sets, the choice to subset or super set some archived data sets, provide the customer with the ability to register as a user, provide for billing and accounting of orders where appropriate and easy access to on-line help as well as a help desk (human) contact. The physical location of the archived data will be transparent to these customers. |

### 4.4.2 Architecture Characteristics

| Technology Category | Requirements |
|---|---|
| | |
| Application Software | -inventory interface<br>-media management interface<br>- Web gateway<br>- CLASS Customer Access Toolkit (CAT) |
| Process Management Interface | The process manager will be responsible to gather the customer information and direct it to the appropriate CLASS component. |

### 4.4.3 Architecture Interfaces

| Component | Requirement |
|---|---|
| Data Retrieval | two-way information transactions |
| Data Delivery | two-way information transactions |

## 4.5 Data Retrieval/Repackaging

A CLASS data retrieval facility is a 24x7 operation. It will provide archived data sets, subsets of archived data sets, and super sets of archived data sets to all requesting applications. The physical location of the archived data will be transparent to these applications, and may be either in processor cache, on-line digital storage, or long term digital media. If the site operates as a disaster backup, it will automatically retrieve data as requested by the remote site and make that data available via the Data Distribution facility (see 2.6.1).

### 4.5.1 Functional Characteristics

| Process | Description |
|---|---|
| Retrieval | Data will be automatically accessed from cache, from network attached or network accessible (on-line) storage media, or from near-line, “long term” archive storage. Requests for complete data sets will be satisfied by copying (virtually or physically) to customer accessible, on-line storage. Data to be subset or super set will be copied (virtually or physically) to network attached or network accessible on-line storage. |
| Remote Recovery | If data are unavailable from local storage, backup volumes will be retrieved automatically from the remote disaster backup site, transferred to local, network attached on-line storage for subsequent “re-hosting” to local long term storage. Customer requested data will then be processed as in Process 1. Retrieval. |
| Subset/Super set Processing | Process received from the Customer Access facility are used to spawn automated processes which either extract portions of the retrieved data sets or merge multiple data sets. |
| Distribution Notification | The processes which perform data distribution will be automatically notified when data are available for further processing. This includes both data for customers and data for disaster recovery at a remote site. |

### 4.5.2 Architecture Characteristics

| Technology Category | Requirements |
|---|---|
| Telecommunications | No addition to Data Storage |
| Processors | No addition to Data Ingest |
| On-line Storage | No addition to Data Ingest |
| Long term Storage | No addition to Data Storage |
| Application Software | -inventory interface<br>-media management interface |
| Process Management Interface | No addition to Data Storage |

### 4.5.3 Architecture Interfaces

| Component | Requirement |
|---|---|
| Customer Access | two-way information transactions |
| | . |

## 4.6 Data Distribution

A CLASS data distribution facility will be a 24x7 operation. This operation is time critical in terms of NESDIS customer satisfaction. Distribution targets are established and guaranteed such that selected data are made available within *n* hours of request where *n* is determined through implementation of NESDIS data distribution policies and procedures. The facility may be manned or unmanned at any point in time, however the operation is automatically and continuously monitored such that any system (hardware/software/telecommunications) or data problems which require human intervention have appropriate maintenance personnel available. If the CLASS site serves as a disaster backup for another site, it will automatically distribute data requested via telecommunications links to the requesting site. Disaster recovery requests will have targeted data availability at *n'* hours of request, where *n'* is mutually agreed upon between the two facilities.

The facility will provide all external (to the site) requested data sets, subsets, and super sets via on-line digital storage which will be physically (or virtually) segregated from all other CLASS processing components, i.e., “outside the firewall”. Data will also be provided for off-line on physical media. Internally requested data sets will be stored on network attached or network accessible on-line storage for further NESDIS processing or reprocessing. Massive amounts of digital storage are characteristic of the retrieval facility.

### 4.6.1 Functional Characteristics

| Process | Description |
|---|---|
| Staging/Access | Data will be resident on online digital storage from the Retrieval facility. For external requests data will be physically or virtually transferred to customer accessible on-line storage areas or written to computer compatible media for off-line delivery. |
| Customer Interface Notification | The processes which perform customer access will be automatically notified when data are available for further processing, if appropriate. |
| Reprocessing Notification | The processes which perform data reprocessing will be automatically notified when data are available for further processing, if appropriate. |

### 4.6.2 Architecture Characteristics

| Technology Category | Requirements |
|---|---|
| Telecommunications | No addition to Data Ingest |
| Processors | No addition to Data Ingest |
| On-line Storage | Dedicated and shared high density digital storage matched to a projection of maximum requested data volume. |
| Long term Storage | No addition to Data Ingest |
| Application Software | - inventory interface<br>- media management interface |
| Process Management Interface | No addition to Data Ingest |

### 4.6.3 Architecture Interfaces

| Component | Requirement |
|---|---|
| Customer Access | two-way information transactions |

| | |
|---|---|
| External Architecture Data Ingest | Process interactions with all external (to CLASS) electronic data recipients, including protocol handshakes, two-way information transactions (messaging), and data throughput |

# 5 Conceptual Views of CLASS Architecture

## 5.1 Data Management Perspective

The following sections provide an aspect of the CLASS architecture from a data and information processing viewpoint.

### 5.1.1 Data Ingest

Data will enter CLASS as: automated, periodic electronic transmissions from suppliers; aperiodic, scheduled electronic transmissions; or scheduled and unscheduled delivery via computer compatible media. Standard data formats, structures and media will be known to the system, i.e., defined in advance, delineated in the CLASS data catalog, and able to have catalog/inventory information extracted from a combination of naming convention and actual data scanning. Acceptable data content (volume, duration, spatial/temporal limits, etc.) will also be known to CLASS. Ingested data will be stored on transitory, on-line devices.

Data will be manipulated through a controlled, iterative process which determines acceptability (to CLASS ingest criteria) and extracts initial CLASS catalog information. This information is made immediately available to all other CLASS functional components via the catalog. Data Quality Assessment and Data Delivery processes will be notified that incoming data is available for further processing. Other CLASS components will be prevented from data access, though they are able to use catalog information for their unique requirements.

Data which fail the initial ingest acceptance criteria will be rejected for further processing, and marked as such in the catalog. An iterative process between CLASS and the data provider will continue until either acceptable data are delivered, or attempts are terminated.

For legacy systems which are only partially CLASS-compliant, information is passed to their unique catalog/inventory systems for further processing. Successfully ingested data are made accessible to these systems as needed.

All activities/processes will be logged to a CLASS management information catalog.

### 5.1.2 Data Quality Assessment

Data will be analyzed for adherence to CLASS guidelines for long term storage. These guidelines will include structure, format, completeness, and non-duplicity. For some lower volume data sets, these guidelines may also include the scientific quality determined. Where data fail to meet the "non-science" CLASS standards, the Data Ingest process is notified and the iterations between Ingest and supplier are undertaken. Data failing to meet science standards will be made available to non-CLASS systems for further analysis, processing, and interaction with supplier(s)/scientist. The CLASS catalog will not be impeded by the scientific assessment, but updated to indicate the assessment of the data and the status (accepted, rejected, or pending) for long term storage and customer servicing.

All other CLASS components will be notified, via the catalog of all accepted data sets available for further processing. Acceptable data remain on transitory, on-line storage, but is now available for access by Data Storage, Data Delivery, and Customer Access.

All activities/processes will be logged to a CLASS management information catalog.

### 5.1.3 Data Storage

The Storage Management System (SMS) will use the data catalog to determine what ingested data sets are available for migration or replication on long term, off-line (or near-line) storage. These data sets will be copied, though will remain on transitory storage for a time determined on a data type basis, i.e. some data sets like AVHRR may remain for 3 days, other smaller sets may remain for weeks, months, or years depending upon the resources available. Long term storage will include DVD, magnetic tape, magnetic cartridges, et. al., and will be determined by the volume of the data sets and the resources available at each facility. The type of media is of no consequence to CLASS, other than the ability to automatically locate and retrieve the data.

All data residing in CLASS will have long term disaster backup copies at remote physical (but logically within) CLASS facilities. The remote Data Storage facility will periodically scan the CLASS catalog for appropriate data sets, access those data sets across the CLASS intranet, copy them to remote long term media, and notify the local SMS of completion.

The local Data Storage component may operate as a backup for other sites. Again, using the CLASS catalog, the local site will scan for appropriate data sets, copy to long term media, and notify the originating site. There is no requirement that remote, disaster backup site maintain data on transitory storage for any length of time beyond what is required to effect the long term storage of that data.

All activities/processes will be logged to a CLASS management information catalog.

### 5.1.4 Data Retrieval

The Data Retrieval component will be actively notified by either the Data Delivery component or a remote Data Retrieval component of a request for full data sets. The SMS will locate the data using CLASS standards for data set naming. Data which reside on transitory media will be immediately available and no further processing will be required. Data available only on long term storage will be retrieved to transitory, on-line media and the requesting component notified of availability. The SMS catalog will be updated to note that data are now available on transitory storage. As with Data Storage, rules governing the duration that data continue to reside on transitory media will apply.

All activities/processes will be logged to a CLASS management information catalog.

### 5.1.5 Data Delivery

The Data Delivery component receives requests for delivery from Customer Access which include tasking for full data sets, subsets, super-sets, and products which may include reformatting or data “co-mingling”. The Data Retrieval component will receive requests from the local Customer Access component for CLASS data which may not reside at the local site. Data Delivery serves as a middleman, parsing those requests to either the local or remote CLASS Data Retrieval components or to CLASS-compliant sites. Physical data location will be determined through the CLASS catalog, and any remote sites directed to provide the data sets as they would for a request at their particular facility. Notification of availability will be forwarded by the remote site. For fully compliant CLASS sites, full data sets, subsets, and super-sets would be provided to customers directly by that site.

For data sub-setting, super-setting or data reformatting the Data Delivery component of the requesting CLASS site will provide all process execution, accessing any remote data across the CLASS intranet and providing the resultant product on local, transitory storage. CLASS-compliant sites will provide only full data sets and will not be responsible for generating CLASS products or services beyond data availability. The resultant product will be stored on transitory media and the Customer Access component notified of its location and attributes.

All activities/processes will be logged to a CLASS management information catalog.

### 5.1.6 Customer Access

All data residing in CLASS will be available to the Customer Access component. Additionally, data sets residing on external systems which are CLASS-compliant will have all or a portion available. Logically, this will appear as a single large array data facility. Physically, the data will reside on either transitory or long term storage located at the several CLASS and CLASS-compliant sites.

The physically distributed, logically singular CLASS catalog will maintain the location of each data set. The Customer Access component will formulate a request for one or more data sets using customer criteria and notify its local Data Retrieval component of the nature of its request (full data set, subset, super-set, product, etc.). As Data Retrieval completes each task assigned, the Customer Access component will be notified and the customer either presented with the data or product via the Web or notified of its availability to download or be shipped.

All activities/processes will be logged to a CLASS management information catalog.

## 5.2 Customer Perspective

There will be three distinct categories of CLASS customers. Each will be served equally well by the CLASS Customer Access component with tailored access pathways, information gathering and display methodologies, and levels of service provided. For each category, CLASS may appear different, though the underlying processes and intelligence will be the same.

### 5.2.1 The Casual Customer.

The casual or novice customer includes first time CLASS customers, customers who want to understand what CLASS is and what services are provided, and customers whose information needs far out way their data needs. The general public, including K-12, are included. Individuals who may have used global Web search engines looking for such diverse entities as "satellites", "oceans", "sunspots", or "climate" will be included. The casual customer will not be expected to purchase data and information and need not register or become "known" to CLASS. This customer will have little impact on the technological resources required to operate and maintain CLASS.

For these customers, the CLASS architecture will provide:

- a "one system view", i.e., unlike many multi system Web information bases, the customer will not have the impression that he is visiting multiple sites, with multiple visualization styles, and multiple methodologies for acquiring information;

- an intelligent information base such that general questions about environmental data and data products and their usefulness may be answered interactively;

- an intelligent information base such that keywords or key phrases may be used to access information about data, data products, and information within and external (but known) to CLASS;

- an intelligent information base such that specific questions about CLASS data, data products, and information may be answered interactively, including such diverse issues as scope and scale of data holdings, what data sets may be logically merged, what data sets may be subset or super set, what products are free and what products require cost recovery, and data access and delivery options available;

- an intelligent information base such that the CLASS Customer Access Toolkit (CAT) is explained, including visualization tools, data analysis tools, information reporting tools, and data reformatting/restructuring tools;

- a virtually seamless portal to freely accessible data and information and the CLASS CAT for manipulation of such freely accessible data and information, as well as the porting of any client-side tools to the customer site;

- a means for establishing an electronic or off-line dialogue with CLASS facilities, including customer help services; and

- a means for becoming a basic customer, including an explanation of cost recovery policies and practices, and for registering their identification with CLASS, such as USER ID and PASSWORD.

### 5.2.2 The Basic Customer

The basic customer includes those individuals who have a need for environmental data and information that is beyond the scope of the casual customer. This customer will have the option of bypassing the more generalized information provided to the casual customer, though such information will always be an accessible option. The basic customer includes both those who are interested only in freely accessible data and information and those who have the option and means of paying the nominal cost of data reproduction. This customer may select data to be downloaded within the guidelines established for data volume or have data provided and delivered on computer compatible media. For server-side portions of CAT, these customers will have a measurable impact on CLASS operational resources.

For these customers, the CLASS architecture will provide:

- a means for accessing the advanced data and product access and delivery portion of CLASS, such as USER ID and PASSWORD verification;

- a means for optionally providing all the functionality of the casual customer;

- a means for accessing more detailed information about data sets, including quality assessments, product algorithms, data supplier information, etc.;

- a means for saving and restoring selection criteria, including data sets required, spatial and temporal criteria, data quality and other attributes required;

- a means for establishing and terminating subscription services;

- a virtually seamless portal to the CLASS CAT for manipulation of all CLASS data and data products;

- a means for ordering data and data products for electronic or off-line delivery;

- a means for paying for those data and product selections which require cost recovery, including interactive credit card or credit voucher authorization and execution;

- a means for tracking the status of current orders and the history of past orders;

- a means for accessing data sets which have been previously ordered, but not as yet downloaded to the customer's site;

- a means for request for and authorization of power customer status, if required;

- a portal, though not seamless, to legacy on-line systems which are not (as yet) CLASS compliant.

### 5.2.3 The Power Customer.

The power customer requires immediate access to data selection, data analysis, and data delivery tools. She will not require any "hand holding" beyond the provision of a gateway into CLASS. This customer will have previously established payment methodologies for data acquisition and delivery where required. The power customer will have the option of becoming a "push" customer, i.e, to have data selections ported directly to the customer site. This customer category will include CLASS facility scientists and engineers who require access to data to satisfy data management requirements (research, data mining, tool evaluation, software testing, data migrations, etc.) which are beyond the scope of CLASS operations. The power customer will also include facilities such as the NPP SDS who require access to data and metadata held within CLASS, but not necessarily accessible by other customers (proprietary). This customer category will have a major impact on the operational resources of CLASS.

The CLASS architecture will provide:

- a means for accessing the advanced data and product access and delivery portion of CLASS, such as USER ID and PASSWORD verification;

- a means for optionally providing **all** of the functionality of both casual and basic customers;

- a means for establishing an addressable gateway between CLASS server(s) and the customer client host for either subscription or customer selection "push" delivery;

- a means for ordering electronically delivered data which exceeds standard CLASS limits for both volume and/or data type (proprietary data);

- a means for ordering "bulk" off-line data delivery;

- client-side tools for complex data selection and manipulation which can be exercised using a point-and-click technique; and

- an environment for operating and maintaining customer created server-side tools to be used only by individual customers or individual sites (e.g., the SDS, the NNDC, PMEL).

## 6.0 CLASS Hardware Architecture

At this time there does not appear to be any technology hurdles in hardware, system software or telecommunications that would keep CLASS from meeting the stated objectives. Technology trends for the past 5-8 years has shown many breakthroughs which have driven costs, per unit, down even though costs per whole system remains the same or at a modest increase. Depending on the marketability of a technology breakthrough, vendors initial prices (to recuperate research and development costs) can be high until quantities are sold or competitors provide similar technologies. The upcoming years (estimated to 2008) show even more upgrades to current technology and expect even more breakthroughs, based on prototyping in labs. Although it is impossible to predict the future (and most vendors would rather not) as the writing of this document, the trends for unit costs is expected to continue downward.

## 6.1 Assumptions used in estimating H/W needs:

(1) Each data category (EOS, NEXRAD, etc.) is treated separately. These estimates assume no sharing of hardware. There could be some savings in processor, disks, and tape libraries if some of the data categories shared hardware. However, there seems to be a tendency to NOT share hardware so that one data category is not "starved" by another satellite instrument when storing and/or retrieving data.

(2) Processing needs are for data store/retrieve handling only. Significant QA or software compression may require additional processors. This analysis assumes a minimum of one control processor node and one data mover node per data category. It is assumed that one data mover (IBM Winterhawk 375Mhz class processor) is required per 100 TB of annual data ingest. (The 100 TB/Yr per mover node is a result of the IBM Study of the HDSS for NCDC, July 14, 2000)

(3) On-line disk capacity should be equal to twice the amount of primary data ingested per day, plus the expected amount of backup data ingested per day.

(4) The compressibility percentages, in Data Estimates Used, were taken from the IBM Study of the HDSS for NCDC, July 14, 2000.

(5) The percent of data that is stored as off-site backup is assumed to be 100%.

(6) The estimated annual growth rates are constant for each year

## 6.2 Data Estimates Used

| Data Category | Est. Annual Growth (TB) | % Compressible | % Duplicated |
|---|---|---|---|
| POES (DMSP) | 13.0 | 20 | 100 |
| GOES | 17.5 | 50 | 100 |
| NPP | 1,000.0 | 20 | 100 |
| NPOESS | 2,000.0 | 20 | 100 |
| NEXRAD | 61.5 | 80 | 100 |
| NOMADS | 95.0 | 20 | 100 |
| NOS Side Scan | 126.0 | 50 | 100 |
| EOS | 3,000.0 | 20 | 100 |
| MetOp | 520.0 | 20 | 100 |

## 6.3 Estimated Processing Needs for the Data Categories

| Data Categories | No. Nodes 2CPUs per | Approx. Cost* |
|---|---|---|
| POES (DMSP) | 2 | $ 80,000 |
| GOES | 2 | $ 80,000 |
| NPP | 11 | $ 440,000 |
| NPOESS | 21 | $ 840,000 |
| NEXRAD | 2 | $ 80,000 |
| NOMADS | 2 | $ 80,000 |
| NOS Side Scan | 3 | $ 120,000 |
| EOS | 31 | $1,240,000 |
| MetOp | 7 | $ 380,000 |

*Cost information uses the cost trends that are in Section 7.1.

## 6.4 Estimated On-Line Disk Storage for the Data Categories

| Data Categories | Est. Disk (GB) | Approx. Cost* |
|---|---|---|
| POES (DMSP) | 107 | $ 7,479 |
| GOES | 144 | $ 10,068 |
| NPP | 8,219 | $ 575,342 |
| NPOESS | 16,438 | $ 1,150,684 |
| NEXRAD | 505 | $ 35,383 |
| NOMADS | 781 | $ 54,657 |
| NOS Side Scan | 1,036 | $ 72,493 |
| EOS | 24,658 | $ 1,726,027 |
| MetOp | 4,274 | $ 299,178 |

*Cost information uses the cost trends that show disk storage will run between $.06-$.08 per megabyte, $.07 was used in this estimate. NOTE: these figures must be adjusted for RAID mirroring, parity drives, spares and proper compression when used for procurement reasons.

## 6.5 Estimated Near-Line Tape Library Storage for the Data Categories, based on estimated starting year under CLASS

| Data Categories | Estimated Start Year | Year 1 of Data Start, in TB (compressed) | Approx. Cost* |
|---|---|---|---|
| POES (DMSP) | 2002 | 10 | $11,100 |
| GOES | 2002 | 9 | $9,990 |
| NPP | 2005 | 800 | $312,000 |
| NPOESS | 2010 | 1,600 | $400,000 |
| NEXRAD | 2003 | 12 | $7,200 |
| NOMADS | 2005 | 76 | $29,640 |
| NOS Side Scan | 2005 | 63 | $24,570 |
| EOS | 2004 | 2,400 | $946,000 |
| MetOp | 2006 | 416 | $162,240 |

*Cost information uses the cost trends that are in Section 7.2, based on the start year.

## 6.6 Estimated Cumulative Back-up Storage for the Data Categories

| Data Categories | Year 1 of Data Start, in TB | Year 2 of Data Start in TB | Year 3 of Data Start in TB | Year 4 of Data Start in TB | Year 5 of Data Start in TB |
|---|---|---|---|---|---|
| POES (DMSP) | 10 | 21 | 31 | 42 | 52 |
| GOES | 9 | 18 | 26 | 35 | 44 |
| NPP | 800 | 1,600 | 2,400 | 3,200 | 4,000 |
| NPOESS | 1,600 | 3,200 | 4,800 | 6,400 | 8,000 |
| NEXRAD | 12 | 25 | 37 | 49 | 62 |
| NOMADS | 76 | 152 | 228 | 304 | 380 |
| NOS Side Scan | 63 | 126 | 189 | 252 | 315 |
| EOS | 2,400 | 4,800 | 7,200 | 9,600 | 12,000 |
| MetOp | 416 | 832 | 1,248 | 1,664 | 2,080 |

## 7.0 Cost Trends

The following costs are estimates as of today. As pointed out in the next section 8.0, it is recommended that a new study be done to update these estimates, and the technology direction on a yearly basis.

## 7.1 Central Processing Unit (CPU)

Based on the IBM Power4 technology, however, the trend is representative of CPU future enhancements:

| Year | 2001 | 2003 | 2005 | 2008 |
|---|---|---|---|---|
| SP System | Regatta H | Regatta H+ | Armada+ | Blue Light* |
| Clock Speed | 1.2Ghz | 1.8Ghz | 3.0Ghz | >3.0Ghz |

| | | | | |
|---|---|---|---|---|
| Peak Perf./cpu | 4.8GF/cpu** | 7.2GF/cpu | 12GF/cpu | >12GF/cpu |
| Peak GF/node | 154GF | 230GF | 768GF | >768gf |
| Max Mem./node | 256GB | 512GB | 1TB | 8TB |
| Max CPU/node | 32 | 32 | 64 | 256 |
| Est. Cost per CPU | $20,000 | $20,000 | $20,000 | $20,000 |
| Cost per Mflop*** | $4.10 | $2.70 | $1.60 | <$.10 |

* Blue Light is a research project that hopes to deliver a system with a peak performance of 256TeraFlops with 8TB of memory and 1Tb/s of system bandwidth with a price performance of $100 per Gigaflop in the 2006 time frame.

** Gigaflop per CPU – one billion floating point operations per second per CPU

***megaflops – one million floating point operations per second

## 7.2 Tape Cartridge

| Year | 2001 | 2003 | 2005 | 2008 |
|---|---|---|---|---|
| LTO Capacity | 100GB | 200GB | 400GB | 800GB |
| LTO Cost per cart. | $111 | $120 | $150 | $200 |
| LTO Cost per GB | $1.11 | $.60 | $.39 | $.25 |
| 3590E Capacity | 40GB | 80GB | 160GB | 320GB |
| 3590E Cost per cart. | $55 | $60 | $75 | $85 |
| 3590E Cost per GB | $1.38 | $.75 | $.47 | $.27 |

The first three generations of LTO are expected to be backward compatible, the fourth generation, in 2008 is not expected to be backward compatible.

## 7.3 Cartridge Library (Robotic)

| Year | 2001 | 2003 | 2005 | 2008 |
|---|---|---|---|---|
| ADIC K10 (2500 slots) | $428,000 | $428,000 | $428,000 | $428,000 |
| ADIC AML/2 (4800) | $490,000 | $490,000 | $490,000 | $490,000 |
| Expansion (4800) | $260,000 | $260,000 | $260,000 | $260,000 |
| Capacity increases K10* | 250TB | 500TB | 1000TB | 2000TB |
| Cap. increases AML* | 480TB | 960TB | 1920TB | 3840TB |
| K10 cost per TB | $1,712 | $856 | $428 | $214 |
| AML cost per TB | $1,020 | $510 | $255 | $127 |

*Based on NO compression

## 7.4 Telecommunications

| Year | 2001 | 2003 | 2005 | 2008 |
|---|---|---|---|---|
| OC3 Per Mile | $55 | $46 | $38 | $26 |
| OC12 Per Mile | $165 | $129 | $107 | $88 |

## 7.5 Initial CLASS H/W Costs

Assuming CLASS will be built to initially ingest POES data, the following hardware costs are projected starting FY02:

| | | |
|---|---|---|
| Processing system (5 nodes) | $400,000 | |
| Disk space (110 GB) | $ 8,000 | |
| Library (K10) | $428,000 | |
| Drives (8) | $100,000 | |
| Storage Management (HPSS) | $325,000 | |
| Media (53TB) | $ 58,830 | (includes current archive) |
| Total | $1,319,830 | |

## 7.6 Continuing H/W Cost Projections

Assuming GOES is added in FY03:

| | | |
|---|---|---|
| Processing Upgrades | $ 80,000 | |
| Disk space (144 GB) | $ 10,000 | |
| Drives (4) | $ 50,000 | |
| Media (282TB) | $313,020 | (includes current archive) |
| Total | $453,020 | |

Assuming NEXRAD is added in FY04:

| | | |
|---|---|---|
| Processing Upgrades | $80,000 | |
| Disk space (500 GB) | $36,000 | |
| Library (AML/2) | $490,000 | |
| Drives (8) | $100,000 | |
| Media (302TB) | $181,200 | (includes current archive) |
| Total | $887,200 | |

Assuming EOS is added in FY04:

| | |
|---|---|
| Processing Upgrades | $1,240,000 |

| | | |
|---|---|---|
| Disk space (2500 GB) | $1,700,000 | |
| Library (AML/2 Expansion) | $260,000 | |
| Drives (8) | $100,000 | |
| Media (2400TB) | $1,440,000 | (includes current archive) |
| Total | $4,740,000 | |

Assuming NPP is added in FY05:

| | |
|---|---|
| Processing Upgrades | $440,000 |
| Disk space (4200 GB) | $300,000 |
| Drives (4) | $ 50,000 |
| Media (800TB) | $312,000 |
| Total | $1,102,000 |

Assuming MetOp is added in FY06:

| | |
|---|---|
| Processing Upgrades | $280,000 |
| Disk space (4200 GB) | $300,000 |
| Drives (4) | $ 50,000 |
| Media (416TB) | $162,240 |
| Total | $792,240 |

Assuming NPOESS is added in FY10:

| | | |
|---|---|---|
| Processing Upgrades | $840,000 | |
| Disk space (16,500 GB) | $1,100,000 | |
| Library (AML/2 Expansion) | $260,000 | |
| Drives (8) | $100,000 | |
| Media (1600TB) | $400,000 | (includes current archive) |
| Total | $2,700,000 | |

## 8.0 Recommendations

The following recommendations are based on the overall requirements study for archive and access:

**8.1** Based on the current hardware in NESDIS and the industry study (attached for reference), the following are recommended:

(1) IBM RS/6000 technology (SP Winterhawk (Power3 @ 375Mhz) and Nighthawk (Power3 @ 625Mhz) nodes). Current technology (good through 2004-2005) at a cost of about $15K per CPU. Refresh to Power4 technology at a cost of about $20K per CPU in the 2005 time frame.

(2) ADIC K10 Near-line Robotic Storage (LTO and, when released, 3590 drives) Up to 2.5PB with 100GB cartridges (uncompressed). LTO drives cost $12,500 each the K10 holds 2500 cartridges, max 16 drives. K10 cost about $428K. An alternative to the K10 would be the ADIC AML/2.

(3) Linear Tape Open (LTO) cartridges, cost about $111 each today.

(4) Storage Management Systems (SMS), although the HPSS is felt to be the leader in this field for large array data sets we feel the recommended CLASS concept will allow each site to use it's own (affordable) SMS. That is the CLASS system should be able to interface with any of the current SMS's in NESDIS, and not just one. Each CLASS site can download the customized CLASS application for their mission & SMS.

*(It should be noted that the industry study documents (by Amdahl), see reference material provided, provides for a total CLASS architecture solution without re-use of current systems with expansion through 2005)*

**8.2** Although the hardware mentioned in section 8.1 are what is recommended today, they may not be the "right" system for tomorrow. Therefore, a short term study (4-5 weeks) should be done annually, under the direction of the ITAT, or by the ITAT, to evaluate current technologies. This study could be the basis for the first study, and each study can be based on the previous one. The point is to see if the direction NESDIS has chosen is still the one it wants to follow. Plus the "estimated" future costs can be validated or updated. This annual study should be budgeted for each year with justification the very fast pace of technology trends, especially in mass storage.

**8.3** NESDIS, through the ITAT, should develop a refreshment plan for hardware and media. (ex., PC life is expected to be 5 years, refresh every 3 years (or 1/3 of PC's each year, UNIX system expected life is 7 years, refresh every 5 years (or 1/5 each year), Tape library life is expected to be 10 years, refresh every 8 years, etc) This plan should also be reflected in the annual budget.

A migration plan should also be made based on age of legacy media to current technology trends (like LTO). This should be an automated process for tape media, to take place within the near-line storage via the storage management system, if possible. Media costs should also be a annual budget item.

**8.4** One CLASS software system, updated on a recurring basis to add more data sets, keep up with technology trends, fix problems, add enhancements and remain portable. The system should be under a strong configuration management discipline that has version control for each release of the software. The software should be distributed to all CLASS sites via a secure FTP site.

**8.5** A CLASS network storage topology should be planned and implemented which follows emerging standards and technologies supported by the National Storage Industry Consortium (NSIC) working group on Network-Attached Storage Devices (NASD). In particular, emphasis should be placed on distributed, network addressable storage devices which relieve the burden of I/O processing on system computing resources and share that burden among intelligent storage systems. Such approaches as rapid deployment of new storage devices (plug-and-play technologies); migration of storage-specific functions out of the operating systems; high-level compatibility among diverse vendors; automated capacity planning and capacity sharing; dynamic data staging across network resources; fastpathing access; and storage-device to storage-device (peer-to-peer) transfers should be explored and utilized in designing, implementing, and operating CLASS.

**8.6** One legacy data set selected from at least two of the four NESDIS facilities (OSDPD, NCDC, NGDC, NODC) for CLASS prototyping. (e.g., logical choices for collaborative access and data discovery:

1. CLASS Portal prototype designed using above data sets
2. CLASS Configuration Management Plan developed
3. CLASS Configuration Control Board (ITAT?) Assigned
4. Legacy data sets documented under CM
5. Process defined for addition of future data sets to CLASS under CM
6. Portal design and subsequent development, testing, implementation integrated through CM Plan
7. Prototype (V0.1) *beta* tested with live customers
8. GAA designed using CLASS CM procedures with goal of adding GOES access through prototype Portal
9. Portal V0.1 through V0.n implemented for bug fixes, required enhancements, etc.
10. *n* legacy data sets determined for first CLASS enhancement
11. Portal 1.0 implemented with CLASS access to the GAA (if available) and the legacy data sets chosen
12. A plan to integrate all other data set types (future and past)